

k -center Clustering under Perturbation Resilience *

Maria-Florina Balcan

Nika Haghtalab

Colin White

Abstract

The k -center problem is a canonical and long-studied facility location and clustering problem with many applications in both its symmetric and asymmetric forms. Both versions of the problem have tight approximation factors on worst case instances: a 2-approximation for symmetric k -center and an $O(\log^*(k))$ -approximation for the asymmetric version. Therefore to improve on these ratios, one must go beyond the worst case.

In this work, we take this approach and provide strong positive results both for the asymmetric and symmetric k -center problems under a very natural input stability (promise) condition called α -*perturbation resilience* [11], which states that the optimal solution does not change under any α -factor perturbation to the input distances. We show that by assuming 2-perturbation resilience, the exact solution for the asymmetric k -center problem can be found in polynomial time. To our knowledge, this is the first problem that is hard to approximate to any constant factor in the worst case, yet can be optimally solved in polynomial time under perturbation resilience for a constant value of α . Furthermore, we prove our result is tight by showing symmetric k -center under $(2 - \epsilon)$ -perturbation resilience is hard unless $NP = RP$. This is the first tight result for any problem under perturbation resilience, i.e., this is the first time the exact value of α for which the problem switches from being NP-hard to efficiently computable has been found.

Our results illustrate a surprising relationship between symmetric and asymmetric k -center instances under perturbation resilience. Unlike approximation ratio, for which symmetric k -center is easily solved to a factor of 2 but asymmetric k -center cannot be approximated to any constant factor, both symmetric and asymmetric k -center can be solved optimally under resilience to 2-perturbations.

*Authors' addresses: {ninamf, nhaghtal, crwhite}@cs.cmu.edu. This work was supported in part by NSF grants CCF-0953192, CCF-1451177, CCF-1422910, a Sloan Research Fellowship, a Microsoft Research Faculty Fellowship, a Google Research Award, an IBM Ph.D. fellowship, and a National Defense Science & Engineering Graduate (NDSEG) fellowship.

1 Introduction

Overview: Traditionally, the theory of algorithms has focused on the analysis of worst-case instances. While this approach has led to many elegant algorithms and strong lower bounds, it tends to be overly pessimistic of an algorithm’s performance on the most typical instances of a problem. A recent line of work in the algorithms community, the so called *beyond worst case analysis* of algorithms, considers the question of designing algorithms for instances that satisfy some natural structural properties and has given rise to strong positive results [3, 4, 6, 20, 23, 24, 28]. One of the most appealing properties that has been proposed in this space is the stability of the solution to small changes in the input. Bilu and Linial [11] formalized this property in the notion of α -perturbation resilience, which states that the optimal solution does not change under any α -factor perturbation to the input distances.

A large body of work has sought to exploit the power of perturbation resilience in problems such as center-based clustering [4, 9, 11, 25], finding Nash equilibria in game theoretic problems [8], and the traveling salesman problem [26]. These works are focused on providing positive results for exactly solving the corresponding optimization problem under perturbation resilient instances, for example, $1+\sqrt{2}$ -perturbation resilience for center based clustering, and $O(\sqrt{\log n \log \log n})$ -perturbation resilience for max-cut. In this work we continue this line of work and provide a tight result for the canonical and long-studied k -center clustering problem, thereby completely quantifying the power of perturbation resilience for this problem. We show that $\alpha = 2$ is the moment where the problem switches from NP-hard to efficiently computable – specifically, we show that by assuming 2-perturbation resilience, the exact solution for the k -center problem can be found in polynomial time; we also show that k -center under $(2 - \epsilon)$ -perturbation resilience cannot be solved in polynomial time unless $NP = RP$. Our results apply to both symmetric and asymmetric k -center, illustrating a surprising relationship between symmetric and asymmetric k -center instances under perturbation resilience. Unlike approximation ratio, for which symmetric k -center is easily solved to a factor of 2 but asymmetric k -center cannot be approximated to any constant factor, both symmetric and asymmetric k -center can be solved optimally under resilience to 2-perturbations. Overall, this is the first tight result quantifying the power of perturbation resilience for a canonical combinatorial optimization problem.

Our Results: The k -center problem is a canonical and long-studied clustering problem with many applications to facility location, data clustering, image classification, and information retrieval [12, 13, 14, 15, 17, 30]. For example, it can be used to solve the problem of placing k fire stations spaced throughout a city to minimize the maximum time for a fire truck to reach any location, given the pairwise travel times between important locations in the city. In the symmetric k -center problem the distances are assumed to be symmetric, while in the asymmetric k -center problem they are not; however in both cases they satisfy the triangle inequality. Formally, given a set of n points S , a distance function $d : S \times S \rightarrow \mathbb{R}^+$ satisfying the triangle inequality (and symmetry in the symmetric case), and an integer k , our goal is to find k centers $\{c_1, \dots, c_k\}$ to minimize $\max_{p \in S} \min_i d(c_i, p)$.

Both forms of k -center admit tight approximation bounds. For symmetric k -center, several 2- approximation algorithms have been found starting in the mid 1980s (e.g., [17, 21]). This is the best possible approximation factor by a simple reduction from set cover. On the other hand, the asymmetric k -center problem is a prototypical problem where the best known approximation is superconstant and is matched by a lower bound. For the asymmetric k -center problem, an $O(\log^*(n))$ -approximation algorithm was found by Vishwanathan [30], and later improved to $O(\log^*(k))$ by Archer [1]. This approximation ratio was shown to be asymptotically tight by the work of Chuzhoy et al. [15], which built upon a sequence of papers establishing the hardness of approximating d -uniform hypergraph covering (culminating in [16]).

In this work we consider both symmetric and asymmetric k -center under perturbation resilience and give tight results for both forms. In addition, we consider more robust and weaker variants of perturbation resilience, and give strong results for these problems as well. A summary of our results and techniques used to achieve them are as follows:

1. *Efficient algorithm for symmetric and asymmetric k -center under 2-perturbation resilience.* This directly improves over the result of Balcan and Liang [9] for symmetric k -center under $1 + \sqrt{2}$ -perturbation resilience. We show that *any* α -approximation algorithm returns the optimal solution for an α -perturbation resilient instance, thus showing there exists an optimal algorithm for symmetric k -center under 2-perturbation resilience. For the asymmetric result, we first construct a “symmetrized set” by only considering points that demonstrate a rough symmetry. Then we prove strong structural results about the symmetrized set which motivates a novel algorithm for detecting clusters locally.
2. *Hardness of symmetric k -center under $(2 - \epsilon)$ -perturbation resilience.* This shows that our perturbation-resilience results are tight for both symmetric and asymmetric k -center. For this hardness result, we use a reduction from a variant of perfect dominating set. To show that this variant is itself hard, we construct a chain of parsimonious reductions (reductions which conserve the number of solutions) starting from 3-dimensional matching to perfect dominating set.
3. *Efficient algorithms for symmetric and asymmetric k -center under $(3, \epsilon)$ -perturbation resilience.* A clustering instance satisfies (α, ϵ) -perturbation resilience if $\leq \epsilon n$ points switch clusters under any α -perturbation. We assume the optimal clusters are of size $> 2\epsilon n$ (the problem is NP-hard without this assumption). We show that if any single point p is close to an optimal cluster other than its own, then $k - 1$ centers achieve the optimal score under a carefully constructed 3-perturbation. Any other point we add to the set of centers must create a clustering that is ϵ -close to OPT , and we show all of these sets cannot simultaneously be consistent with one another, thus causing a contradiction. A key concept in our analysis is defining the notion of a *cluster-capturing center*, which allows us to reason about which points can capture a cluster when its center is removed.
4. *Efficient algorithm for any center-based clustering objective under weak center proximity.* Weak center proximity asks that each point be closer to its own center than to any point from any other cluster, but note that it allows a cluster center to be closer to points from different clusters than to its own. Thus it is not at all obvious whether efficient optimal clustering is possible in such a setting. We present a novel linkage-based algorithm that is able to do so. It works by iteratively running single linkage as a subroutine until all clusters are balanced, and then removing all but the very last link.

The novelty of our results are manifold. First, our work is the first to provide a tight perturbation resilience result, thereby, painting the complete picture for k -center under perturbation resilience. Second, this is the first result where a problem is not approximable to any constant in the worst-case, but can be optimally solved under resilience to small constant perturbations. Third, we are the first to consider an asymmetric problem under stability. Our results here illustrate a stark contrast between worst-case analysis and analysis of algorithms under stability. Unlike approximation ratio, for which symmetric k -center is easily solved to a factor of 2 but asymmetric k -center cannot be approximated to any constant factor, both symmetric and asymmetric k -center can be solved optimally under the same constant level of resilience.

Perturbation resilience has a natural interpretation for both symmetric and asymmetric k -center: it can be viewed as a stability condition in the presence of uncertainties involved in measurements. For example, small fluctuations in the travel time between a fire station and locations in the city, which are caused by different levels of traffic at different times of day, should not drastically affect the optimal placement of fire stations. Furthermore, perturbation resilience can be viewed as a condition on an instance under which the optimal solution satisfies a form of privacy. For instance, if the actions of no individuals (such as how they drive to work, or the amount of network traffic they are using in a network application) can affect the overall state of the problem drastically then the individual’s actions cannot be detected by looking at the optimal solution.

1.1 Related Work

Perturbation Resilience There has been a recent substantial interest on providing algorithms that circumvent worst case hardness results on stable, realistic instances. Bilu and Linial defined Perturbation Resilience [11] and showed algorithms that found the optimal solutions for $(1 + \epsilon)$ -perturbation resilient instances of metric and dense Max-Cut, and $\Omega(\sqrt{n})$ -perturbation resilient instances for Max-Cut in general. The latter bound was improved by Markarychev et al. [25], who showed the standard SDP relaxation is integral for $\Omega(\sqrt{\log n \log \log n})$ -perturbation resilient instances. This result is similar in flavor to our Theorems 4 and 11 in that algorithms used in practice give the optimal solution assuming perturbation resilience. They show that finding an algorithm for $o(\sqrt{\log n})$ -perturbation resilience would improve the best known approximation for Sparsest Cut, but such an algorithm must either output an optimal solution, or certify the instance is not stable (our Theorem 10 applies to any algorithm that is required to produce the optimal solution). Finally, they showed an algorithm to find the optimal solution for 4-perturbation resilient instances of Minimum Multiway Cut. Awasthi et al. [4] studied α -perturbation resilience under center-based clustering objectives (which includes k -median, k -means, and k -center clustering, as well as other objectives), showing an algorithm to optimally solve 3-perturbation resilient instances. Balcan and Liang [9] improved this result, finding an algorithm to optimally solve center-based objectives under $(1 + \sqrt{2})$ -perturbation resilience. They also studied (α, ϵ) -perturbation resilience, showing an algorithm that finds near-optimal solutions for k -median under $(4, \epsilon)$ -perturbation resilience, assuming a lower bound on the size of the optimal clusters. Recent work has applied perturbation resilience to other settings to obtain better than worst-case approximation guarantees, including finding Nash equilibria in game theoretic problems [8] and the travelling salesman problem [26].

Now we describe results for a related stability assumption called approximation stability, which is strictly stronger than perturbation resilience.

Approximation Stability Balcan et al. [6] defined (α, ϵ) -approximation stability, a related, stronger notion than perturbation resilience, and showed algorithms that returned near-optimal solutions for k -median and k -means when $\alpha > 1$, and min-sum when $\alpha > 2$. Gupta et al. [19] gave an alternative algorithm for finding near-optimal solutions for k -median under approximation stability as part of their work on finding structure in triangle-dense graphs. The result for the min-sum objective was later improved by Balcan and Braverman [7] to $\alpha > 1$, while also doing better when there is no lower bound on the size of the optimal clusters. Voevodski et al. [31] gave an algorithm for optimally clustering protein sequences using the min-sum objective under approximation stability, which empirically compares favorably to well-established clustering algorithms.

Other Stability Assumptions There has also been work on other types of stability assumptions for clustering. Ovstrosky et al. [27] studied a separation condition in which the k -means cost of a clustering instance is much lower than the $(k - 1)$ -means cost. They showed how to efficiently cluster these instances using the Lloyd heuristic with a random farthest traversal seeding. Kumar and Kannan [23] studied a condition in which the projection of any point onto the line between its cluster center to any other cluster center is a large additive factor closer to its own center than the other center. They showed that using this assumption one can efficiently cluster the data. These results were later improved along several axes by Awasthi and Sheffet [5]. Many other works have shown strong positive results on instances satisfying beyond worst case natural notions of stability on problems ranging from clustering to data privacy to social networks to topic modeling [2, 3, 19, 20, 23, 24, 28].

2 Preliminaries

We define a clustering instance as (S, d) , where S is a set of n points and $d : S \times S \rightarrow \mathbb{R}_{\geq 0}$ is a distance function. In the k -center problem, the goal is to find a set of points $\mathbf{p} = \{p_1, \dots, p_k\} \subseteq S$ called *centers* such

that the maximum distance from any point to its closest center is minimized. More formally, in the k -center problem, given a Voronoi partition $\mathcal{P} = \{P_1, \dots, P_k\}$ induced by a set of centers $\mathbf{p} = \{p_1, \dots, p_k\}$ (where for all $1 \leq i \leq k$, $p_i \in P_i$), we define the cost of \mathcal{P} by $\Phi(\mathcal{P}) = \max_{i \in [k]} \max_{v \in P_i} d(p_i, v)$. We indicate by \mathcal{OPT} the clustering $\{C_1, \dots, C_k\}$ with minimum cost, we denote the optimal centers $\{c_1, \dots, c_k\}$, and we denote the optimal cost $\Phi(\mathcal{OPT})$ by r^* , the maximum cluster radius.

We study the k -center clustering of instance (S, d) under two types of distance functions, *symmetric* and *asymmetric*. A symmetric distance function is a metric. An asymmetric distance function satisfies all the properties of a metric space, except for symmetry. That is, it may be the case that for some $p, q \in S$, $d(p, q) \neq d(q, p)$. Note that the k -center objective function for asymmetric instances is the same as the symmetric case, the maximum distance *from the center to the points*, where the order now matters.

We consider *perturbation resilience*, a notion of stability introduced by Bilu & Linial [11]. Perturbation resilience implies that the optimal clustering does not change under small perturbations of the distance measure. Formally, d' is called an α -perturbation of distance function d , if for all $p, q \in S$, $d(p, q) \leq d'(p, q) \leq \alpha d(p, q)$.¹ Perturbation resilience is defined formally as follows.

Definition 1. A clustering instance (S, d) satisfies α -perturbation resilience for k -center, if for any α -perturbation d' of d , the optimal k -center clustering under d' is unique and equal to \mathcal{OPT} .

Note that the optimal centers may change, but the Voronoi partition C_1, \dots, C_k induced by them must stay the same. We do *not* assume that d' is a metric.² We also consider a more robust variant of α -perturbation resilience, called (α, ϵ) -perturbation resilience, that allows a small change in the optimal clustering when distances are perturbed. To this end, we say that two clusterings \mathcal{C} and \mathcal{C}' are ϵ -close, if only an ϵ -fraction of the input points are clustered differently in the two clusterings, i.e., $\min_{\sigma} \sum_{i=1}^k |C_i \setminus C'_{\sigma(i)}| \leq \epsilon n$, where σ is a permutation on $[k]$. Formally,

Definition 2. A clustering instance (S, d) satisfies (α, ϵ) -perturbation resilience for k -center, if for any α -perturbation d' of d , any optimal k -center clustering \mathcal{C}' under d' is ϵ -close to \mathcal{OPT} .

We use ϵ -far to denote two clusters which are not ϵ -close. We also discuss the strictly stronger notion of approximation stability [6], which requires *any* α -approximation (not just a Voronoi partition) to be ϵ -close to \mathcal{OPT} . This is formally defined in Section 3.3. In Section 5, we define *center-based* objectives [9], a more general class of clustering functions which includes objective functions such as k -center, k -median, and k -means. Throughout this work, we use $B_r(c)$ to denote a ball of radius r centered at point c . Also for a point p and a set D , $d(p, D)$ denotes the distance from p to the farthest point in D .

3 2-perturbation resilience

In this section, we provide efficient algorithms for finding \mathcal{OPT} for symmetric and asymmetric instances of k -center under 2-perturbation resilience. Our result directly improves on the result of Balcan and Liang for symmetric k -center under $(1 + \sqrt{2})$ -perturbation resilience [9]. We also show that it is NP-hard to recover \mathcal{OPT} even in the symmetric k -center instance under $(2 - \epsilon)$ -approximation stability. As an immediate consequence, our results are tight for both perturbation resilience and approximation stability, for symmetric and asymmetric k -center instances. This is the first problem for which the exact value of perturbation resilience is found ($\alpha = 2$), where the problem switches from efficiently computable to NP-hard.

In the remainder of this section, first we show that any α -approximation algorithm returns the optimal solution for α -perturbation resilient instances. An immediate consequence is an algorithm for symmetric

¹WLOG, we only consider perturbations in which the distances increases because we can scale the distances to simulate decreasing distances.

²This is well-justified, as the data may be gathered from heuristics or an average of measurements.

k -center under 2-perturbation resilience. Then we provide a novel algorithm for asymmetric k -center under 2-perturbation resilience.

3.1 Approximation algorithms under perturbation resilience

The following lemma allows us to reason about a specific type of α -perturbation we construct. This lemma will be important throughout the analysis in this section and in Section 4.

Lemma 3. *For all $\alpha \geq 1$, given an α -perturbation d' of d with the following property: for all p, q , if $d(p, q) \geq r^*$ then $d'(p, q) \geq \alpha r^*$. Then the optimal cost under d' is αr^* .*

Proof. Clearly the optimal cost under d' cannot be greater than αr^* , since d' is an α -perturbation. Suppose there exists a set of centers c'_1, \dots, c'_k under d' that achieves a cost $< \alpha r^*$. Then for all i and all $p \in C'_i$, $d'(c'_i, p) < \alpha r^*$. But then by assumption, $d(c'_i, p) < r^*$. This implies that c'_1, \dots, c'_k achieve an optimal cost $< r^*$ under d , which is a contradiction. \square

The following theorem will imply that any α -approximation algorithm for k -center will return the optimal solution on clustering instances that are α -perturbation resilient.

Theorem 4. *Given a clustering instance (S, d) satisfying α -perturbation resilience for asymmetric k -center. Given a set C of k centers which is an α -approximation, i.e., $\forall p \in S, \exists c \in C$ s.t. $d(c, p) \leq \alpha r^*$. Then the Voronoi partition induced by C is the optimal clustering.*

Proof. For a point $p \in S$, let $c(p) := \operatorname{argmin}_{c \in C} d(c, p)$, the closest center in C to p . The idea is to construct an α -perturbation in which C is the optimal solution by increasing all distances except between p and $c(p)$, for all p . Then the theorem will follow by using the definition of perturbation resilience.

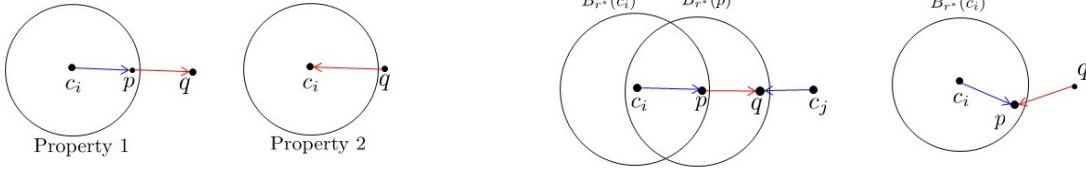
By assumption, $\forall p \in S, d(c(p), p) \leq \alpha r^*$. Create a perturbation d' as follows. Increase all distances by a factor of α , except for all $p \in S$, set $d'(c(p), p) = \min(\alpha d(c(p), p), \alpha r^*)$. Then no distances were increased by more than a factor of α . And since we had that $d(c(p), p) \leq \alpha r^*$, no distances decrease either. Therefore, d' is an α -perturbation of d . By Lemma 3, the optimal cost for d' is αr^* . Also, C achieves cost $\leq \alpha r^*$ by construction, so C is an optimal set of centers under d' . Then by α -perturbation resilience, the Voronoi partition induced by C under d' is the optimal clustering.

Finally, we show the Voronoi partition of C under d is the same as the Voronoi partition of C under d' . Given $p \in S$ whose closest point in C is $c(p)$ under d , then under d' , all distances from p to $C \setminus \{c(p)\}$ increased by exactly α , and $d(p, c(p))$ increased by $\leq \alpha$. Therefore, the closest point in C to p under d' is still $c(p)$. \square

An immediate consequence is that we have exact algorithms for symmetric k -center under 2-perturbation resilience, and asymmetric k -center under $O(\log^*(k))$ -perturbation resilience. Now we show it is possible to substantially improve the latter result.

3.2 Asymmetric k -center algorithm

One of the challenges involved in dealing with asymmetric k -center instances is the fact that even though for all $p \in C_i, d(c_i, p) \leq r^*, d(p, c_i)$ might be arbitrarily large. Such points for which $d(p, c_i) \gg r^*$ pose a challenge to the structure of the clusters, as they can be very close to points or even centers of other clusters. To deal with this challenge, we first define a set of “good” points, A , such that $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$. Intuitively speaking, these points behave similarly to a set of points with symmetric distances up to a distance r^* . To explore this, we define a desirable property of A with respect to the optimal clustering.



(a) Properties of 2-perturbation resilience

(b) Demonstrating the correctness of Algorithm 1

Figure 1: Properties of a 2-perturbation resilient instance of asymmetric k -center that are used for clustering.

Definition 5. A is said to respect the structure of \mathcal{OPT} if (1) $c_i \in A$ for all $i \in [k]$, and (2) for all $p \in S \setminus A$, if $A(p) := \arg \min_{q \in A} d(q, p) \in C_i$, then $p \in C_i$.

For all i , define $C'_i = C_i \cap A$ (which is in fact the optimal clustering of A , although we do not need to prove this). Definition 5 implies that if we can optimally cluster A , then we can optimally cluster the entire instance (formalized in Theorem 8). Thus our goal is to show that A does indeed respect the structure of \mathcal{OPT} , and to show how to return C'_1, \dots, C'_k .

Intuitively, A is similar to a symmetric 2-perturbation resilient clustering instance. However, some structure is no longer there, for instance, a point p may be at distance $\leq 2r^*$ from every point in a different cluster, which is not true for 2-perturbation resilient instances. This implies we cannot simply run a 2-approximation algorithm on the set A , as we did in the previous section. However, we show that the remaining structural properties are sufficient to optimally cluster A . To this end, we define two properties and show how they lead to an algorithm that returns C'_1, \dots, C'_k , and help us prove that A respects the structure of \mathcal{OPT} .

The first of these properties requires each point to be closer to its center than any point in another cluster. That is, *Property (1):* For all $p \in C'_i$ and $q \in C'_j$, $i \neq j$, $d(c_i, p) < d(q, p)$. The second property requires that any point within distance r^* of a cluster center belongs to that cluster. That is, *Property (2):* For all $i \neq j$ and $q \in C_j$, $d(q, c_i) > r^*$ (Figure 1).³

Let us illustrate how these properties allow us to optimally cluster A .⁴ Consider a ball of radius r^* around a center c_i , by Property 2, such a ball exactly captures C'_i . Furthermore, by Property 1, any point in this ball is closer to the center than to points outside of the ball. Is this true for a ball of radius r^* around a general point p ? Not necessarily. If this ball contains a point $q \in C'_j$ from a different cluster, then q will be closer to a point outside the ball than to p (namely, c_j , which is guaranteed to be outside of the ball by Property 2). This allows us to determine that the center of such a ball must not be an optimal center.

This structure motivates our Algorithm 1 for asymmetric k -center under 2-perturbation resilience. At a high level, we start by constructing the set A (which can be done easily in polynomial time). Then we create the set of all balls of radius r^* around all points in A (if r^* is not known, we can use a guess-and-check wrapper). Next, we prune this set by throwing out any ball that contains a point farther from its center than to a point outside the ball. We also throw out any ball that is a subset of another one. Our claim is that the remaining balls are exactly C'_1, \dots, C'_k . Finally, we add the points in $S \setminus A$ to their closest point in A .

Lemma 6. *Properties 1 & 2 hold for asymmetric k -center instances under 2-perturbation resilience.*

Proof sketch. For Property 2, assume that there exists c_i and $q \in C_j$, $i \neq j$, such that $d(q, c_i) \leq r^*$. We construct a 2-perturbation in which q becomes the center for C_i (similar to the previous paragraph). Increase

³ Property (1) first appeared in the work of Awasthi et al. [4], for symmetric clustering instances. A weaker variation of Property (2) was introduced by Balcan and Liang [9], which showed that in $1 + \sqrt{2}$ -perturbation resilient instances for any cluster C_i with radius r_i , $B_{r_i}(c_i) = C_i$. Our Property (2) shows that this is true for a universal radius, r^* , even for 2-perturbation resilient instances, and even for asymmetric instances.

⁴ Other algorithms work, such as single linkage with dynamic programming at the end to find the minimum cost pruning of k clusters. However, our algorithm is able to recognize optimal clusters *locally* (without a complete view of the point set).

Algorithm 1 ASYMMETRIC k -CENTER ALGORITHM UNDER 2-PR

Input: Asymmetric k -center instance (S, d) , r^* (or try all possible candidates).

1. Build set $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$
2. $\forall c \in A$, construct $G_c = B_{r^*}(c)$ (the ball of radius r^* around c).
3. $\forall G_c$, if $\exists p \in G_c, q \notin G_c$ s.t. $d(q, p) < d(c, p)$, then throw out G_c .
4. $\forall p, q$ s.t. $G_p \subseteq G_q$, throw out G_p .
5. $\forall p \notin A$, add p to G_q , where $q = \arg \min_{s \in A} d(s, p)$.

Output: Output the sets G_1, \dots, G_k .

all distances by a factor of 2, except for the distances from q to C_i , which we increase until they reach $2r^*$. By Lemma 3, this 2-perturbation achieves a cost of $2r^*$. However, q is distance $2r^*$ to C_i , so it must replace c_i as an optimal center. Then q and c_j are no longer in the same cluster, causing a contradiction.

The first property was shown to hold for symmetric instances by Awasthi et al. and the same proof can be used for asymmetric instances. We include the proof in Appendix A. \square

Lemma 7. A respects the structure of \mathcal{OPT} .

Proof sketch. First we show that $c_i \in A$ for all $i \in [k]$. Given $c_i, \forall p \in C_i, d(c_i, p) \leq r^*$ by definition of \mathcal{OPT} . $\forall q \notin C_i$, by Property 2, $d(q, c_i) > r^*$. Therefore, for any point $p \in S$, it cannot be the case that $d(p, c_i) \leq r^*$ and $d(c_i, p) > r^*$.

Now we show that for all $p \in S \setminus A$, if $A(p) \in C_i$, then $p \in C_i$. Assume towards contradiction that $\exists p \in C_i \setminus A$ and $q = A(p) \in C_j$ for some $i \neq j$. Similar to the proof of Lemma 22, we will construct a 2-perturbation d' in which q replaces c_j as the center for C_j . We do this by increasing all distances by a factor of 2 except for $d(q, q'), \forall q' \in C_j$, which we increase by a factor of 2 up to $2r^*$. But then, since all centers are in A , and q is the closest point in A to p , it follows that in d' , p switches clusters from C_i to C_j . This completes the proof. \square

Theorem 8. Algorithm 1 returns the exact solution for asymmetric k -center under 2-perturbation resilience.

Proof. First we must show that after step 4, the remaining sets are exactly $C'_1, \dots, C'_k = C_1 \cap A, \dots, C_k \cap A$. We prove this in three steps: the sets G_{c_i} correspond to C'_i , these sets are not thrown out in steps 3 and 4, and all other sets are thrown out in steps 3 and 4. Because of Lemma 22, we can use Properties 1 and 2.

For all $i, G_{c_i} = C'_i$: From Lemma 23, all centers are in A , so G_{c_i} will be created in step 2. For all $p \in C_i, d(c_i, p) \leq r^*$. For all $q \notin C'_i$, then by Property 2, $d(q, c_i) > r^*$ (and since $c_i, q \in A, d(c_i, q) > r^*$ as well). For all i, G_{c_i} is not thrown out in step 3: Given $s \in G_{c_i}$ and $t \notin G_{c_i}$. Then $s \in C'_i$ and $t \in C'_j$ for $j \neq i$. If $d(t, s) < d(c_i, s)$, then we get a contradiction from Property 1. For all non-centers p, G_p is thrown out in step 3 or 4: From the previous paragraph, $G_{c_i} = C'_i$. If $G_p \subseteq G_{c_i}$, then G_p will be thrown out in step 4 (if $G_p = G_{c_i}$, it does not matter which set we keep, so WLOG say that we keep G_{c_i}). Then if G_p is not thrown out in step 4, $\exists s \in G_p \cap C'_j, j \neq i$. If $s = c_j$, then $d(p, c_j) \leq r^*$ and we get a contradiction from Property 2. So, we can assume s is a non-center (and that $c_j \notin G_p$). But $d(c_j, s) < d(p, s)$ from Property 1, and therefore G_p will be thrown out in step 3. Thus, the remaining sets after step 4 are exactly C'_1, \dots, C'_k .

Finally, by Lemma 23, for each $p \in C_i \setminus A, A(p) \in C_i$, so p will be added to G_{c_i} . Therefore, the final output is C_1, \dots, C_k . \square

3.3 Hardness for k -center under $(2 - \epsilon)$ -approximation stability

In this section, we consider approximation stability, introduced by Balcan et al. [6], which is strictly stronger than perturbation resilience. We show that if symmetric k -center under $(2 - \epsilon)$ -approximation stability can be

solved in polynomial time, then $NP = RP$, even under the condition that the optimal clusters are all $\geq \frac{n}{2k}$. Because approximation stability is stronger than perturbation resilience, this result implies k -center under $(2 - \epsilon)$ -perturbation resilience is hard as well. Similarly, symmetric k -center is a special case of asymmetric k -center, so we get the same hardness results for asymmetric k -center. This proves that Theorem 8 is tight.

Approximation stability requires constant approximations to the optimal cost to differ from \mathcal{OPT} by at most an ϵ -fraction of the points.

Definition 9. A clustering instance (S, d) satisfies (α, ϵ) -approximation stability for k -center, if for any partition \mathcal{C}' with objective value r' (not necessarily a Voronoi partition), if $r' \leq \alpha r^*$, then \mathcal{C}' is ϵ -close to \mathcal{OPT} .

It is not hard to see that (α, ϵ) -approximation stability implies (α, ϵ) -perturbation resilience, as the optimal clustering under any α -perturbation costs at most αr^* under the original distance function, d . So, a violating instance of (α, ϵ) -perturbation resilience induces a partition which costs $\leq \alpha r^*$ and is ϵ -far from \mathcal{OPT} , and therefore is not (α, ϵ) -approximation stable.

Theorem 10. There is no polytime algorithm for finding the optimal k -center clustering under $(2 - \epsilon)$ -approximation stability, even when assuming all optimal clusters are size $\geq \frac{n}{2k}$, unless $NP = RP$.

We show a reduction from a special case of Dominating Set which we call Unambiguous-Balanced-Perfect Dominating Set. A reduction from Perfect Dominating Set (Dominating Set with the additional constraint that for all dominating sets of size $\leq k$, each vertex is hit by exactly one dominator) to the problem of clustering under $(2 - \epsilon)$ -center proximity was shown in [10] (α -center proximity is the property that for all $p \in C_i$ and $j \neq i$, $\alpha d(c_i, p) < d(c_j, p)$), and it follows from α -perturbation resilience). Our contribution is to show that Perfect Dominating Set remains hard under two additional conditions. First, in the case of a YES instance, each dominator must hit at least $\frac{n}{2k}$ vertices (which translates to clusters having size at least $\frac{n}{2k}$ as well). Second, we are promised that there is at most one dominating set of size $\leq k$ (which is required for establishing approximation stability for the resulting clustering instance). The details are provided in Appendix B.

4 Robust perturbation resilience

In this section, we consider (α, ϵ) -perturbation resilience. We show that under $(3, \epsilon)$ -perturbation resilience, there is an algorithm that recovers \mathcal{OPT} for symmetric k -center, and an algorithm that returns a solution that is ϵ -close to \mathcal{OPT} for asymmetric k -center. For both of these results, we assume a lower bound on the size of the optimal clusters, $|C_i| > 2\epsilon n$ for all $i \in [k]$. We show the lower bound on cluster sizes is necessary; in its absence, the problem becomes NP-hard for all values of $\alpha \geq 1$ and $\epsilon > 0$. The theorems in this section require a careful reasoning about sets of centers under different perturbations that cannot all simultaneously be valid.

4.1 Symmetric k -center

We show that for any $(3, \epsilon)$ -perturbation resilient k -center instance such that $|C_i| > 2\epsilon n$ for all $i \in [k]$, \mathcal{OPT} can be found by simply thresholding the input graph using distance r^* and outputting the connected components. A nice feature of our result is that the Single Linkage algorithm, a fast algorithm widely used in practice, is sufficient to optimally cluster these instances.

Theorem 11. Given a $(3, \epsilon)$ -perturbation resilient k -center instance (S, d) where all optimal clusters are $> \max(2\epsilon n, 3)$. Then the optimal clusters in \mathcal{OPT} are exactly the connected components of the threshold graph G_{r^*} of the input distances.

Proof idea. Since each optimal cluster center is distance r^* from all points in its cluster, it suffices to show that any two points in different clusters are at least r^* apart from each other. Assume on the contrary that there exist $p \in C_i$ and $q \in C_j$, $i \neq j$, such that $d(p, q) \leq r^*$. First we find a set of $k + 2$ points and a 3-perturbation d' , such that every size k subset of the points are optimal centers under d' . Then we show how this leads to a contradiction under $(3, \epsilon)$ -perturbation resilience.

From our assumption, p is distance $\leq 3r^*$ from every point in $C_i \cup C_j$ (by the triangle inequality). Under a 3-perturbation in which all distances are blown up by a factor of 3 except $d(p, C_i \cup C_j)$, then replacing c_i and c_j with p would still give us a set of $k - 1$ centers that achieve the optimal score. But, *would this contradict $(3, \epsilon)$ -perturbation resilience?* Indeed, not! Perturbation resilience requires exactly k *distinct* centers.⁵ The key challenge is to pick a final “dummy” center to guarantee that the Voronoi partition is ϵ -far from \mathcal{OPT} . The dummy center might “accidentally” be the closest center for almost all points in C_i or C_j . Even worse, it might be the case that the new center sets off a chain reaction in which it becomes center to a cluster C_x , and c_x becomes center to C_j , which would also result in a partition that is not ϵ -far from \mathcal{OPT} .

To deal with the chain reactions, we crucially introduce the notion of a *cluster capturing center* (CCC). c_x is a CCC for C_y , if for all but ϵn points $p \in C_y$, $d(c_x, p) \leq r^*$ and for all $i \neq x, y$, $d(c_x, p) < d(c_i, p)$. Intuitively, a CCC exists if and only if c_x is a valid center for C_y when c_y is taken out of the set of optimal centers (i.e., a chain reaction will occur). We argue that if a CCC does not exist then every dummy center we pick must be close to either C_i or C_j , since there are no chain reactions. If there does exist a CCC c_x for C_y , then we cannot reason about what happens to the dummy centers under our d' . However, we can define a new d'' by increasing all distances except $d(c_x, C_y)$, which allows us to take c_y out of the set of optimal centers, and then any dummy center must be close to C_x or C_y . There are no chain reactions because *we already know c_x is the best center for C_y among the original optimal centers*. Thus, whether or not there exists a CCC, we can find $k + 2$ points close to the entire dataset by picking points from both C_i and C_j (resp. C_x and C_y).

Because of the assumption that all clusters are size $> 2\epsilon n$, for every 3-perturbation there must be a bijection between clusters and centers, where the center is closest to the majority of points in the corresponding cluster. We show that all size k subsets of the $k + 2$ points cannot simultaneously admit bijections that are consistent with one another.

Formal analysis. We start out with a simple implication from the assumption that $|C_i| > 2\epsilon n$ for all i .

Fact 12. *Given a clustering instance which is (α, ϵ) -perturbation resilient for $\alpha \geq 1$, and all optimal clusters have size $> 2\epsilon n$. Then for any α -perturbation, for any set of optimal centers c'_1, \dots, c'_k , for each optimal cluster C_i , there must be a unique c'_i which is the center for more than half of the points in C_i under d' .*

Now we formally define a CCC.

Definition 13. *A center c_i is a first-order cluster-capturing center (CCC) for C_j if for all $x \neq j$, for more than half of the points $p \in C_j$, $d(c_i, p) < d(c_x, p)$ and $d(c_i, p) \leq r^*$. c_i is a second-order cluster-capturing center (CCC2) for C_j if there exists a c_l such that for all $x \neq j, l$, for more than half of points $p \in C_j$, $d(c_i, p) < d(c_x, p)$ and $d(c_l, p) \leq r^*$ (see Figure 2a).*

Each cluster C_j can have at most one CCC c_i because c_i is closer than any other center to more than half of C_j . Every CCC is a CCC2, since the former is a stronger condition. However, it is possible for a cluster to have multiple CCC2's.⁶ We needed to define CCC2 for the following reason. Assuming there exist $p \in C_i$

⁵ This distinction is well-motivated; if for some application, the best k -center solution is to put two centers at the same location, then we could achieve the exact same solution with $k - 1$ centers. That implies we should have been running k' -center for $k' = k - 1$ instead of k .

⁶ In fact, a cluster can have at most three CCC2's, but this is not relevant to our analysis.

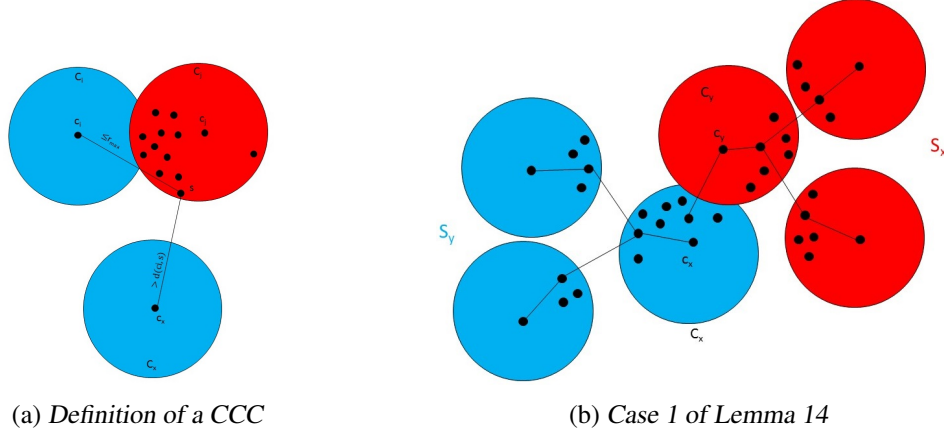


Figure 2: (a) c_i is a CCC for C_j . c_x is a CCC2 for C_j . (b) Case 1 of Lemma 14: there exists a CCC2 c_x . All distances denoted with a black edge are length $\leq r^*$.

and $q \in C_j$ which are close, and we replace c_i and c_j with p in the set of centers. Maybe C_i has a CCC, but it is c_j . This is not relevant to our analysis, since c_j was taken out of the set of centers. However, if we know that c_x is a CCC2 (it is the best center for C_i , disregarding c_j), then we know that c_x will be the best center for C_i after replacing c_i and c_j with p . Now we use this definition to show that if two points from different clusters are close, then all points are close together in some sense. We include the proof in Appendix D.

Lemma 14. *Given a clustering instance satisfying $(3, \epsilon)$ -perturbation resilience and all optimal clusters are size $> 2\epsilon n$. Assume there are two points from different clusters which are $\leq r^*$ apart from each other. Then there exists a partition $S_x \cup S_y$ of S such that for all $p, q \in S_x$, $d(p, q) \leq 2r^*$ and similarly for all $p, q \in S_y$, $d(p, q) \leq 2r^*$*

So far, we have shown that by just assuming two points from different clusters are close, then many points are very close together, in some sense. Now we will show that such an instance cannot be stable under $(3, \epsilon)$ -perturbation resilience.

One implication of the last lemma is that there must exist a set of $k + 2$ points that are collectively close to every point in the dataset (we will formalize this soon). So, there is one 3-perturbation for which any k of these points can be an optimal set of centers. But it is not possible that all $\binom{k+2}{k}$ of these sets of centers simultaneously create partitions that are ϵ -close. This idea is first presented in a more general format. We include the proof in Appendix D.

Lemma 15. *Given a set U of k elements and a set C (disjoint from U) of $k + 2$ elements, and each $u \in U$ ranks all elements in C without ties. It is not possible that for all sets $C' \subseteq C$, $|C'| = k$, each $c \in C'$ is ranked highest by exactly one $u \in U$.*

Looking ahead, each element in U will correspond to a cluster, and each element in C will correspond to a point that becomes an optimal center under a d' we construct.

Note 16. *The lemma will still hold even if each U only ranks its top three elements in C , and all the rest are tied in fourth. This is because for each C' , U only needs to express its top-ranked element. So there can be a C' in which $u \in U$'s first- and second-highest ranked elements are not in C' , in which u needs to specify its third-highest ranked element, but it does not need to uniquely rank any other elements.*

We are almost ready to prove our main lemma. First, we state a fact that helps us apply the setting of Lemma 15 to a clustering instance. We would like to have a table such that each row i is an ordering of all

points based on their distance to C_i . Then we know that in any optimal set of centers, the point that is the center for C_i is the highest point in the ranking. The following fact shows this ranking is well-defined.

Fact 17. *Given a clustering instance (S, d) satisfying (α, ϵ) -perturbation resilience such that all optimal clusters are size $> 2\epsilon n$. Given an α -perturbation with optimal score x . For all i , there exists a ranking of S such that for all sets $|C| = k$ with score x , the unique center in C which is closest to all but ϵn points in C_i , is the point in S ranked highest.*

Proof. Assume the lemma is false. Then there exists a cluster C_i , two points p and q , and two sets of centers $p, q \in C$ and $p, q \in C'$ which achieve score x , but p is the center for C_i in C while q is the center for C_i in C' . Then p is closer than all other points in C to all but ϵn points in C_i . Similarly, q is closer than all other points in C' to all but ϵn points in C_i . Since $|C_i| > 2\epsilon n$, this causes a contradiction. \square

Now using the previous two lemmas, we can prove Theorem 11.

Proof of Theorem 11. It suffices to prove that any two points from different clusters are at distance $> r^*$ from each other. Assume towards contradiction that this is not true. Then by Lemma 14, there exists a partition S_x, S_y of S such that for all $p, q \in S_x$, $d(p, q) \leq 2r^*$ and similarly for S_y . Furthermore, for all c_i , if $\exists p \in C_i \cap S_x$, then all points in S_x are distance $3r^*$ to c_i , otherwise all points in S_y are distance $3r^*$ to c_i .

Construct a set C of size $k + 2$ as follows. Pick at least 3 points from S_x and 3 points from S_y . The rest of the points may be arbitrary. S_x and S_y must contain at least the clusters C_x and C_y , respectively, so here we assume optimal clusters are size ≥ 3 . Set C has the property that for all $p \in S$, at least 3 points in C are distance $\leq 3r^*$ to p . We will use Lemma 15 to show a contradiction.

First we construct a d' in which any size k subset of C is a valid set of centers:

$$d'(p, q) = \begin{cases} 3r^* & \text{if } p \in C \text{ and } r^* \leq d(p, q) \leq 3r^* \\ 3d(p, q) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation by construction. By Lemma 3, the optimal score under d' is $3r^*$. And given any set $C' \subseteq C$, $|C'| = k$, for all $p \in S$, there must exist at least one point $q \in C'$ such that $d(p, q) \leq 3r^*$. Therefore C' is an optimal set of centers.

Define the set $U = \{C_1, \dots, C_k\}$. From Fact 17, for all i , there must be a unique $c_1^{(i)} \in C$ such that for all other points $p \in C$, $c_1^{(i)}$ is closer than p to the majority of points in C_i . Similarly, there must be a unique point $c_2^{(i)} \in C \setminus \{c_1^{(i)}\}$ that is closer to the majority of points in C_i , or else every C' without $c_1^{(i)}$ would have a contradiction by Fact 12. Finally, when we pick the $C' = C \setminus \{c_1^{(i)}, c_2^{(i)}\}$, there must be a unique $c_3^{(i)}$ closer to the majority of points in C_i , for the same reason. Let all $C_i \in U$ define its ranking as $c_1^{(i)}, c_2^{(i)}, c_3^{(i)}$, and all the rest are tied in fourth. (Because of Note 16, it is okay that we have ties for fourth). Now we can use Lemma 15 on U, C . Then there exists a C' such that a $c \in C'$ is ranked highest by at least two elements $C_i, C_j \in U$. Then by definition of the rankings, c is the closest point in C' to the majority of points in C_i and C_j . But then (since each cluster size is $> 2\epsilon n$) the optimal set of centers C' under d' is not ϵ -close to OPT . \square

Note that Theorem 10 implies $(2 - \delta, \epsilon)$ -perturbation resilient k -center is hard for $\delta > 0$, even when the optimal clusters are large. Therefore, the value of α we achieve is within one of optimal.

4.2 Lower bound on cluster sizes

Before moving to the asymmetric case, we show that the lower bound on the cluster sizes in Theorem 11 is necessary. Without this lower bound, clustering becomes hard, even assuming (α, ϵ) -perturbation resilience for any α and ϵ . This reduction follows from k -center (the details appear in Appendix D).

Theorem 18. For all $\alpha \geq 1$ and $\epsilon > 0$, finding the optimal solution for k -center under (α, ϵ) -perturbation resilience is NP-hard.

4.3 $(3, \epsilon)$ -perturbation resilience for asymmetric k -center

In the asymmetric case, we consider the definition of the symmetric set A from Section 3, $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$. We might first ask whether A respects the structure of \mathcal{OPT} , as it did under 2-perturbation resilience. Namely, whether *Condition 1*: all centers are in A , and *Condition 2*: $\arg \min_{q \in A} d(q, p) \in C_i \implies p \in C_i$ hold. This is not the case for either condition. We explore to what degree these conditions are violated.

We call a center “bad” if it is not in the set A , i.e., $\exists q \notin C_i$ and $d(q, c_i) \leq r^*$. When a bad center c_i exists, we can take it out of the set of optimal centers, and we can pick an arbitrary dummy center which must be close to C_i or a CCC for C_i . In our symmetric argument, we arrived at a contradiction by showing that two dummy centers which capture the same cluster, must be close by the triangle inequality. This logic breaks down for asymmetric distances. In Appendix D, we show an example of an instance with a bad center that satisfies (α, ϵ) -perturbation resilience. However, it turns out that *no* instance can have more than 6 bad centers under $(3, \epsilon)$ -perturbation resilience, assuming all optimal clusters have size $> 2\epsilon n$. This is our main structural result for this section (Lemma 20). So Condition 1 is satisfied for all but a constant number of centers. However, Condition 2 may not be satisfied for up to ϵn points. Therefore, even if we fully cluster A , we will only get ϵ -close to \mathcal{OPT} .

Every point in A is at distance r^* from its center and distance $2r^*$ from its entire cluster. However, as we mentioned it is possible that 6 centers are not in A (and possibly no points at all from those 6 clusters). This motivates the following algorithm. First, we run a symmetric k -center 2-approximation algorithm on A , for $k - 6 \leq k' \leq k$. For instance, iteratively pick an unmarked point, and mark all points distance $2r^*$ away from it [21]. This gives us a 2-approximation for the centers in A , and thus a 3-approximation for S minus the clusters with no centers in A . Then we brute force search for the remaining ≤ 6 centers to find a 3-approximation for S . Under $(3, \epsilon)$ -perturbation resilience, this 3-approximation must be ϵ -close to \mathcal{OPT} . Now we state the theorem, and the main structural lemma, which shows that at most 6 centers are bad. We give the full proof in Appendix D.

Algorithm 2 $(3, \epsilon)$ -PERTURBATION RESILIENT ASYMMETRIC k -CENTER

Input: Asymmetric k -center instance (S, d) , r^* (or try all possible candidates).

1. Build set $A = \{p \mid \forall q, d(q, p) \leq r^* \implies d(p, q) \leq r^*\}$.
2. Create the threshold graph G with vertices A , and threshold distance r^* . Define a new symmetric k -center instance with A , using the lengths of the paths in the threshold graph.
3. Run a symmetric k -center 2-approximation algorithm on the symmetrized instance. Start with $k' = k - 6$, and increase k' by 1 until the algorithm returns a solution with radius $\leq 2r^*$.
4. Brute force over all size $k - x$ subsets of C and all size x subsets of S for $x \leq 6$, to find a set of size k which is $3r^*$ from all points in S . Denote this set by C' .

Output: Output the Voronoi tiling G_1, \dots, G_k using C' as the centers.

Theorem 19. Algorithm 2 runs in polytime and outputs a clustering that is ϵ -close to \mathcal{OPT} , for $(3, \epsilon)$ -perturbation resilient asymmetric k -center instances s.t. all optimal clusters are size $> 2\epsilon n$.

Lemma 20. Given a $(3, \epsilon)$ -perturbation resilient asymmetric k -center instance such that all optimal clusters are size $> 2\epsilon n$, there are at most 6 centers c_i such that $\exists q \notin C_i$ and $d(q, c_i) \leq r^*$.

Proof idea. Assume the lemma is false. The first step is to construct a set C of $\leq k - 3$ points which are $\leq 3r^*$ from every point in S . Once we find C , we will be able to find 3 dummy centers which contradict $(3, \epsilon)$ -perturbation resilience.

By assumption, there exists a set B , $|B| \geq 7$, of centers c_i such that $\exists q \in C_{i'}, i' \neq i$, such that $d(q, c_i) \leq r^*$. Then $d(c_{i'}, c_i) \leq 2r^*$, and $d(c_{i'}, C_i) \leq 3r^*$. So for all $c_i \in B$, $\{c_l\}_{l=1}^k \setminus \{c_i\}$ is still distance $3r^*$ from every point in S .

To construct C , we carefully pick a subset $B' \subseteq B$ of size ≥ 3 such that no $c_i \in B$ has a $c_{i'}$ also in B' . Then we can set $C = \{c_l\}_{l=1}^k \setminus B'$ and construct a 3-perturbation in which these $k - 3$ centers achieve the optimal score. Now we find a contradiction by showing that not every combination of 3 dummy centers can simultaneously allow ϵ -close clusterings.

If for all $c \in C$, c is the very best center in S for some cluster C_i , then for any choice of dummy centers, c will be the center for C_i and no other cluster, thus it will not affect our analysis; it is as if our instance is size $k' = k - |C|$. Then we can use Lemma 15 to arrive at a contradiction.

When some set of points $C_{bad} \subset C$ are not the best center for a cluster, we cannot use the same lemma, since there are some centers we must include in every subset. However, we can use the Pigeonhole principle to show that at least one $c \in C_{bad}$ is the best center for two clusters, out of all other points in C_{bad} , and we use this to show a contradiction.

5 Weak center proximity

In this section, we consider any center-based objective, not just k -center. A clustering objective function is *center-based* if the solution can be defined by choosing a set of centers $\{c_1, c_2, \dots, c_k\} \subseteq S$, and partitioning S into k clusters $\mathcal{OPT} = \{C_1, C_2, \dots, C_k\}$ by assigning each point to its closest center. Furthermore, 1) The objective value of a given clustering is a weighted sum or maximum of the individual cluster scores; 2) given a proposed single cluster, its score can be computed in polynomial time. k -median, k -means, and k -center are all center-based objectives.

Here, we show a novel algorithm that finds the optimal clustering in instances that satisfy two simple properties: each point is closer to its center than to any point in a different cluster, and we can recognize optimal clusters as soon as they are formed. Formally, we define these properties as

1. **weak center proximity:** For all $p \in C_i$ and $q \in C_j$, $d(c_i, p) < d(p, q)$.
2. **cluster verifiability:** There exists a polytime computable function $f : 2^S \rightarrow \mathbb{R}$ that for $B \subseteq S$, if there is $i \in [k]$ such that $B \subset C_i$, then $f(B) < 0$, and if $B \supseteq C_i$, then $f(B) \geq 0$.

Examples of cluster verifiable instances include any instance where all the optimal clusters are the same size ($f(B) = |B| - \frac{n}{k}$), or where all the optimal clusters have the same k -median/ k -means cost ($f(B) = \Phi(B) - \Phi(\mathcal{OPT})$).

For any center-based objective, weak center proximity is a consequence of 2-perturbation resilience (i.e., Lemma 22), so, our algorithm relies on a much weaker assumption than α -perturbation resilience for $\alpha \geq 2$, when instances are cluster verifiable.

All existing algorithms and analysis for α -perturbation resilience require that for all $p \in C_i$ and $q \in C_j$, $d(c_i, p) < d(c_i, q)$. It is not at all obvious how one can even proceed without such a property, as in its absence, clusters can ‘overlap’. That is, for a cluster with center c_i and radius r , we can not assume that $B_r(c_i)$ only includes points from C_i . Our challenge is then in showing that even in absence of this property, there is still enough structure imposed by the weak center proximity and cluster verifiability to find the optimal clustering efficiently.

Our Algorithm 3 is a novel linkage based procedure. Given a clustering instance (S, d) , we will start with a graph $G = (S, E)$ where $E = \emptyset$. In each round, we do single linkage on the components in G , except we do not merge two components if both are supersets of optimal clusters (indicated by $f(B) \geq 0$). Put the

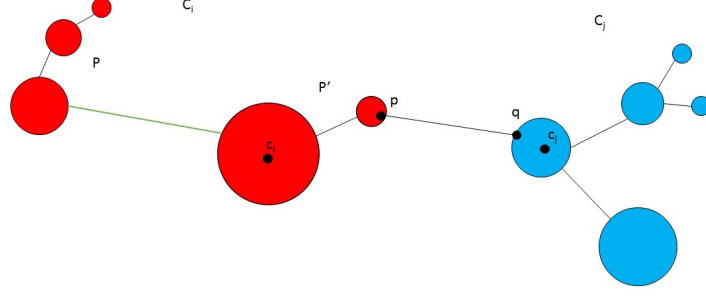


Figure 3: The edge between p and q cannot be the last edge added to A .

single linkage edges from this round in a set A . This will continue until every component is a superset of an optimal cluster. Then we throw away the set A except for the *very last edge* that was added (we include a figure in Appendix E, Figure 3). We will prove this last edge is never between two points from different clusters, so we add that single edge to E and then recur. Here, we present a proof sketch of our main theorem. The details can be found in Appendix E.

Algorithm 3 CLUSTERING UNDER WEAK CENTER PROXIMITY AND CLUSTER VERIFIABILITY

Input: Clustering instance (S, d) , function f , and $k \leq |S|$.

Set $G = (S, E)$ and $E = \emptyset$. While there are more than k components in G , repeat (1) and (2):

1. Set $A = \emptyset$. While there exists a component B in $G' = (S, E \cup A)$ such that $f(B) < 0$, add (p, q) to A , where $d(p, q)$ is minimized such that p and q are in different components in G' and at least one of these components B has $f(B) < 0$.
2. Take the last edge e that was added to A , and put $e \in E$.

Output: Output the components of G .

Theorem 21. *Given a center-based clustering instance satisfying weak center proximity and cluster verifiability, Algorithm 3 outputs OPT in polynomial time.*

Proof Sketch. It suffices to show that step (b) never adds an edge between two points from different clusters. We proceed by induction. Assume it is true up to iteration t of the first while loop. Now assume towards contradiction that in round t , the last edge added to A is between two points $p \in C_i$ and $q \in C_j$, $i \neq j$ (see Figure 3). WLOG, for the component in G' that includes p , called P' , we have $f(P') < 0$, otherwise the merge would not have happened. Furthermore, $c_i \in P'$ by weak center proximity. Then $f(P') < 0$ implies that $C_i \setminus P'$ is nonempty, so call it P . The component(s) in G corresponding to P are strict subsets of C_i , therefore, $f(P) < 0$. So they must merge to another component, and by weak center proximity, the closest component is P' , but this contradicts our assumption that (p, q) was the last edge added to A . \square

6 Conclusions

Our work pushes the understanding of (promise) stability conditions farther in three ways. We are the first to design computationally efficient algorithms to find the optimal clustering under α -perturbation resilience with a constant value of α for a problem that is hard to approximate to any constant factor in the worst case, thereby demonstrating the power of perturbation resilience. Furthermore, we demonstrate the limits of this power by showing the first tight results in this space for both perturbation resilience and approximation stability. Finally, we show a surprising relation between symmetric and asymmetric instances, in that they are equivalent under resilience to 2-perturbations, which is in stark contrast to their widely differing tight approximation factors.

References

- [1] Aaron Archer. Two $o(\log^* k)$ -approximation algorithms for the asymmetric k -center problem. In *Integer Programming and Combinatorial Optimization*, pages 1–14. Springer, 2001.
- [2] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - going beyond SVD. In *53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 1–10, 2012.
- [3] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k -median and k -means clustering. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 309–318, 2010.
- [4] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.
- [5] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. *CoRR*, 2012.
- [6] Maria-Florina Balcan, Avrim Blum, and Anupam Gupta. Approximate clustering without the approximation. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1068–1077, 2009.
- [7] Maria-Florina Balcan and Mark Braverman. Finding low error clusterings. In *COLT*, 2009.
- [8] Maria-Florina Balcan and Mark Braverman. Approximate nash equilibria under stability conditions. Technical report, 2010.
- [9] Maria-Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. In *Automata, Languages, and Programming*, pages 63–74. Springer, 2012.
- [10] Shalev Ben-David and Lev Reyzin. Data stability in clustering: A closer look. In *Algorithmic Learning Theory*, pages 184–198. Springer, 2012.
- [11] Yonatan Bilu and Nathan Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(05):643–660, 2012.
- [12] Fazli Can. Incremental clustering for dynamic information processing. *ACM Transactions on Information Systems (TOIS)*, 11(2):143–164, 1993.
- [13] Fazli Can and ND Drochak. Incremental clustering for dynamic document databases. In *Proceedings of the 1990 Symposium on Applied Computing*, pages 61–67, 1990.
- [14] Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pages 626–635, 1997.
- [15] Julia Chuzhoy, Sudipto Guha, Eran Halperin, Sanjeev Khanna, Guy Kortsarz, Robert Krauthgamer, and Joseph Seffi Naor. Asymmetric k -center is $\log^* n$ -hard to approximate. *Journal of the ACM (JACM)*, 52(4):538–551, 2005.
- [16] Irit Dinur, Venkatesan Guruswami, Subhash Khot, and Oded Regev. A new multilayered pcg and the hardness of hypergraph vertex cover. *SIAM Journal on Computing*, 34(5):1129–1146, 2005.
- [17] Martin E Dyer and Alan M Frieze. A simple heuristic for the p -centre problem. *Operations Research Letters*, 3(6):285–288, 1985.

- [18] Martin E. Dyer and Alan M. Frieze. Planar 3dm is np-complete. *Journal of Algorithms*, 7(2):174–184, 1986.
- [19] Rishi Gupta, Tim Roughgarden, and C Seshadhri. Decompositions of triangle-dense graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 471–482. ACM, 2014.
- [20] Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340, 2013.
- [21] Dorit S Hochbaum and David B Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.
- [22] Harry B Hunt III, Madhav V Marathe, Venkatesh Radhakrishnan, and Richard E Stearns. The complexity of planar counting problems. *SIAM Journal on Computing*, 27(4):1142–1167, 1998.
- [23] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *51st Annual IEEE Symposium on Foundations of Computer Science*, pages 299–308, 2010.
- [24] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time $(1 + \epsilon)$ -approximation algorithm for geometric k-means clustering in any dimensions. In *Proceedings-Annual Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.
- [25] Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Bilu-linial stable instances of max cut and minimum multiway cut. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 890–906. SIAM, 2014.
- [26] Matúš Mihalák, Marcel Schöngens, Rastislav Šrámek, and Peter Widmayer. On the complexity of the metric tsp under stability considerations. In *SOFSEM 2011: Theory and Practice of Computer Science*, pages 382–393. Springer, 2011.
- [27] Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. The effectiveness of lloyd-type methods for the k-means problem. In *47th Annual IEEE Symposium on Foundations of Computer Science*, pages 165–176, 2006.
- [28] Tim Roughgarden. Beyond worst-case analysis. <http://theory.stanford.edu/tim/f14/f14.html>, 2014.
- [29] Leslie G Valiant and Vijay V Vazirani. Np is as easy as detecting unique solutions. *Theoretical Computer Science*, 47:85–93, 1986.
- [30] Sundar Vishwanathan. An $o(\log^* n)$ approximation algorithm for the asymmetric p-center problem. In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1–5, 1996.
- [31] Konstantin Voevodski, Maria-Florina Balcan, Heiko Röglin, Shang-Hua Teng, and Yu Xia. Min-sum clustering of protein sequences with limited distance information. In *Similarity-Based Pattern Recognition*, pages 192–206. Springer, 2011.

A Proofs from Section 3

Lemma 22. *Properties 1 and 2 hold for asymmetric k -center instances satisfying 2-perturbation resilience.*

Proof. Property 1: Assume false, $d(q, p) \leq d(c_i, p)$. The idea will be that since q is in A , it is close to its own center, so we can construct a perturbation in which q replaces its center c_j . Then p will join q 's cluster, causing a contradiction. Construct the following d' :

$$d'(s, t) = \begin{cases} \min(2r^*, 2d(s, t)) & \text{if } s = q, t \in C_j \cup \{p\} \\ 2d(s, t) & \text{otherwise.} \end{cases}$$

This is a 2-perturbation because $d(q, C_j \cup \{p\}) \leq 2r^*$. Then by Lemma 3, the optimal score is $2r^*$. The set of centers $\{c_l\}_{l=1}^k \setminus \{c_j\} \cup \{q\}$ achieves the optimal score, since q is distance $2r^*$ from C_j , and all other clusters have the same center as in \mathcal{OPT} (achieving radius $2r^*$). Then for all c_l , $d'(q, p) \leq d'(c_i, p) \leq d'(c_l, p)$. And since $q \in A$, $d(q, c_j) \leq r^*$ so $d(q, C_j) \leq 2r^*$. Then we can construct a 2-perturbation in which q becomes the center of C_j , and then q is the best center for p , so we have a contradiction.

Property 2: Assume on the contrary that there exists $q \in C_j$, $i \neq j$ such that $d(q, c_i) \leq r^*$. Now we will define a d' in which q can become a center for C_i .

$$d'(s, t) = \begin{cases} \min(2r^*, 2d(s, t)) & \text{if } s = q, t \in C_j \\ 2d(s, t) & \text{otherwise.} \end{cases}$$

This is a 2-perturbation because $d(q, C_j) \leq 2r^*$. Then by Lemma 3, the optimal score is $2r^*$. The set of centers $\{c_l\}_{l=1}^k \setminus \{c_j\} \cup \{p\}$ achieves the optimal score, since p is distance $2r^*$ from C_j , and all other clusters have the same center as in \mathcal{OPT} (achieving radius $2r^*$). But the clustering with centers $\{c_l\}_{l=1}^k \setminus \{c_j\} \cup \{p\}$ is different from \mathcal{OPT} , since (at the very least) p and c_j are in different clusters. This contradicts 2-perturbation resilience. \square

Lemma 23. *A respects the structure of \mathcal{OPT} .*

Proof. From Lemma 22, we can use Property 2 in our analysis. First we show that $c_i \in A$ for all $i \in [k]$. Given c_i , $\forall p \in C_i$, then $d(c_i, p) \leq r^*$ by definition of \mathcal{OPT} . $\forall q \notin C_i$, then by Property 2, $d(q, c_i) > r^*$. It follows that for any point $p \in S$, it cannot be the case that $d(p, c_i) \leq r^*$ and $d(c_i, p) > r^*$. Therefore, $c_i \in A$.

Now we show that for all $p \in S \setminus A$, if $A(p) \in C_i$, then $p \in C_i$. Given $p \in S \setminus A$, let $p \in C_i$ and assume towards contradiction that $q = A(p) \in C_j$ for some $i \neq j$. We will construct a 2-perturbation d' in which q replaces c_j as the center for C_j and p switches from C_i to C_j , causing a contradiction. We construct d' as follows. All distances are increased by a factor of 2 except for $d(q, p)$ and $d(q, q')$ for all $q' \in C_j$. These distances are increased by a factor of 2 up to $2r^*$. Formally,

$$d'(s, t) = \begin{cases} \min(2r^*, 2d(s, t)) & \text{if } s = q, t \in C_j \cup \{p\} \\ 2d(s, t) & \text{otherwise.} \end{cases}$$

This is a 2-perturbation because $d(q, C_j) \leq 2r^*$. Then by Lemma 3, the optimal score is $2r^*$. The set of centers $\{c_l\}_{l=1}^k \setminus \{c_j\} \cup \{q\}$ achieves the optimal score, since q is distance $2r^*$ from C_j , and all other clusters have the same center as in \mathcal{OPT} (achieving radius $2r^*$). But consider the point p . Since all centers are in A and q is the closest point to p in A , then q is the center for p under d' . Therefore, the optimal clustering under d' is different from \mathcal{OPT} , so we have a contradiction. \square

B Proof of Theorem 10

In this section, we prove Theorem 10. The final reduction to k -center under $(2 - \epsilon)$ -approximation stability and large clusters is from a problem we define, called Unambiguous-Balanced-Perfect Dominating Set.

We use four NP-hard problems in a chain of reductions. Here, we define all of these problems up front. Perfect Dominating Set was introduced in [10]. We introduce the “balanced” variants of two existing problems for the first time.

Definition (3-Dimensional Matching (3DM)). *Given three disjoint sets X_1 , X_2 , and X_3 each of size m , and set T such that $t \in T$ is a triple $t = (x_1, x_2, x_3)$, $x_1 \in X_1$, $x_2 \in X_2$, and $x_3 \in X_3$. The problem is to find a set $M \subseteq T$ of size m which exactly hits all the elements in $X_1 \cup X_2 \cup X_3$, in other words, for all pairs $(x_1, x_2, x_3), (y_1, y_2, y_3) \in M$, it is the case that $x_1 \neq y_1$, $x_2 \neq y_2$, and $x_3 \neq y_3$.*

Definition (Balanced-3-Dimensional Matching (B3DM)). *This is the 3DM problem (X_1, X_2, X_3, T) with the additional constraint that $2m \leq |T| \leq 3m$, where $|X_1| = |X_2| = |X_3| = m$.*

Definition (Perfect Dominating Set (PDS)). *Given a graph $G = (V, E)$ and an integer k , the problem is to find a set of vertices $D \subseteq V$ of size k such that for all $v \in V \setminus D$, there exists exactly one $d \in D$ such that $(v, d) \in E$.*

Definition (Balanced-Perfect-Dominating Set (BPDS)). *This is the PDS problem with the additional assumption that if the graph has n vertices and a dominating set of size k exists, then each vertex in the dominating set hits at least $\frac{n}{2k}$ vertices.*

Additionally, each problem has an “Unambiguous” variant, which is the added constraint that the problem has at most one solution. Valiant and Vazirani showed that Unambiguous-3SAT is hard unless $NP = RP$ [29]. To show the Unambiguous version of another problem is hard, one must establish a parsimonious reduction from Unambiguous-3SAT to that problem. A parsimonious reduction is one that conserves the number of solutions. For two problems A and B , we denote $A \leq_{par} B$ to mean there is a reduction from A to B that is parsimonious and polynomial. Note that many reductions which involve 1-to-1 mappings are often trivial to verify parsimony. The problem is when one element in A maps to multiple elements in B . The reductions in this section are all 1-to-1 mappings, and are therefore easy to verify parsimony.

Now we start our argument. Dyer and Freize showed Planar-3DM is NP-hard [18]. While planarity is not important for the purpose of our problems, their reduction from 3SAT has two other nice properties that we crucially depend on. First, the reduction is parsimonious, as pointed out in [22]. Second, given their 3DM instance X_1, X_2, X_3, T , each element in $X_1 \cup X_2 \cup X_3$ appears in either two or three tuples in T . (Dyer and Freize mention this observation just before their Theorem 2.3.) From this, it follows that $2m \leq |T| \leq 3m$, and so they actually showed a stronger result, that B3DM is NP-hard via a parsimonious reduction from 3SAT.

Next, Ben-David and Reyzin showed a reduction from 3DM to PDS [10]. Their reduction maps every element X_1, X_2, X_3, T in the 3DM instance to a vertex in the PDS instance (adding a single extra vertex), so it is easily parsimonious.

We can use the same reduction to show $B3DM \leq_{par} BPDS$. Their reduction maps every element X_1, X_2, X_3, T to a vertex in V , and they add one extra vertex v to V . There is an edge from each element $(x_1, x_2, x_3) \in T$ to the corresponding elements $x_1 \in X_1$, $x_2 \in X_2$, and $x_3 \in X_3$. Furthermore, there is an edge from v to every element in T . In [10], it is shown that if the 3DM instance is a YES instance with matching $M \subseteq T$ then the minimum dominating set is $v \cup M$. Then, this dominating set is size $m + 1$. If we start with B3DM, our graph has $|X_1| + |X_2| + |X_3| + |T| + 1 \leq 6m + 1$ vertices since $|T| \leq 3m$. Given $t \in M$, t hits 3 nodes in the graph, and $\frac{n}{2(m+1)} \leq \frac{6m+1}{2m+2} \leq 3$. Furthermore, v hits $|T| - m \geq 2m - m = m$ nodes, and $\frac{6m+1}{2m+2} \leq m$ when $m \geq 3$. Therefore, the resulting PDS instance is indeed BPDS.

Now we have verified that there exists a chain of parsimonious reductions $3\text{SAT} \leq_{\text{par}} \text{B3DM} \leq_{\text{par}} \text{BPDS}$, so it follows that Unambiguous-BPDS is hard unless $NP = RP$.

At this point, we use the same reduction as in [10] to reduce from Unambiguous-BPDS to k -center clustering under $(2 - \epsilon)$ -approximation stability, where all clusters are size $\geq \frac{n}{2k}$. The difference is that we must verify that the resulting instance is $(2 - \epsilon)$ -approximation stable, which requires the Unambiguity.

Theorem 10. *There is no polynomial time algorithm for finding the optimal k -center clustering under $(2 - \epsilon)$ -approximation stability, even when assuming all optimal clusters are size $\geq \frac{n}{2k}$, unless $NP = RP$.*

Proof. Given $\epsilon > 0$. From the previous discussion, Unambiguous-BPDS is NP-hard unless $NP = RP$. Now we reduce to k -center clustering and show the resulting instance has all cluster sizes $\geq \frac{n}{2k}$ and satisfies $(2 - \epsilon)$ -approximation stability.

Given an instance of Unambiguous-BPDS, for every $v \in V$, create a point $v \in S$ in the clustering instance. For every edge $(u, v) \in E$, let $d(u, v) = 1$, otherwise let $d(u, v) = 2$. Since all distances are either 1 or 2, the triangle inequality is trivially satisfied. Then a k -center solution of cost 1 exists if and only if there exists a dominating set of size k .

Since each vertex in the dominating set hit at least $\frac{n}{2k}$ vertices, the resulting clusters will be size at least $\frac{n}{2k} + 1$.

Additionally, if there exists a dominating set of size k , then the corresponding optimal k -center clustering has cost 1. Because this dominating set is perfect and unique, any other clustering has cost 2. It follows that the k -center instance is $(2 - \epsilon)$ -approximation stable. □

C $(2, \epsilon)$ -Approximation Stability for Symmetric k -center

In this section, we consider k -center clustering under approximation stability (which was defined in Section 3.3).

Since (α, ϵ) -approximation stability is strictly stronger than (α, ϵ) -perturbation resilience, our results from Sections 3 and 4 immediately extend to approximation stability, namely, there are polynomial time algorithms for finding the optimal symmetric or asymmetric k -center clustering under 2-approximation stability, symmetric k -center clustering under $(3, \epsilon)$ -approximation stability, and an ϵ -close algorithm for asymmetric k -center under $(3, \epsilon)$ -approximation stability, where the latter two results assume the optimal clusters are size $> 2\epsilon n$. In this section, we provide an algorithm that finds \mathcal{OPT} for symmetric k -center under $(2, \epsilon)$ -approximation stability, assuming the optimal clusters are size $> \epsilon n$.

A key insight behind our result is that any $p \in C_i$ is at distance $\leq 2r^*$ from all points in C_i (by the triangle inequality). So, any two points in the same cluster have $> \epsilon n$ points in common that are within $2r^*$ of both. On the other hand, if a point $p \in C_i$ were to also be at distance $\leq 2r^*$ from more than ϵn points in other clusters, then replacing p as the center of C_i would lead to a partition that is at most $2r^*$ in cost but is not ϵ -close to \mathcal{OPT} . This contradicts $(2, \epsilon)$ -approximation stability. Therefore, two points from different clusters can only have a small number of points in common that are within $2r^*$ of both of them. Based on this insight, Algorithm 4 proceeds by placing two points p and q , in the same partition iff $|B_{2r^*}(p) \cap B_{2r^*}(q)| > \epsilon n$. We show that this procedure indeed returns a partition that corresponds to \mathcal{OPT} .

Algorithm 4 $(2, \epsilon)$ -APPROXIMATION STABLE k -CENTER FOR LARGE CLUSTERS

Input: Symmetric k -center instance (S, d) , r^* (or try all possible candidates).

1. Define $G = (S, E)$ such that $E = \{(p, q) \mid |B_{2r^*}(p) \cap B_{2r^*}(q)| > \epsilon n\}$.

Output: Connected components of G .

The following theorem formalizes our previous discussion.

Theorem 24. *Given a $(2, \epsilon)$ -approximation stable k -center instance such that for all i , $|C_i| > \epsilon n$, then Algorithm 4 returns \mathcal{OPT} in polynomial time.*

Proof. It suffices to show that for the optimal score r^* , p and q are in the same cluster if and only if $|B_{2r^*}(p) \cap B_{2r^*}(q)| > \epsilon n$.

First we show the forward direction. Assume p and q are in the same cluster C_i . For any pair of points $s_1, s_2 \in C_i$, $d(s_1, s_2) \leq d(s_1, c_i) + d(c_i, s_2) \leq 2r^*$. Therefore $C_i \subseteq B_{2r^*}(p)$ and $C_i \subseteq B_{2r^*}(q)$, so $\epsilon n < |C_i| \leq |B_{2r^*}(p) \cap B_{2r^*}(q)|$.

For the reverse direction, assume on the contrary that there are $p \in C_i$ and $q \in C_j$, $i \neq j$ such that $|B_{2r^*}(p) \cap B_{2r^*}(q)| > \epsilon n$. Take any $\epsilon n + 1$ points from $|B_{2r^*}(p) \cap B_{2r^*}(q)|$ and put them into a set M . Partition M into $M(p)$ and $M(q)$ such that $M(q) = M \cap C_i$ and $M(p) = M \setminus M(q)$. Consider the partition $C_i \cup M(p) \setminus M(q)$ and $C_j \cup M(q) \setminus M(p)$. For all $x \in C_i \cup M(p) \setminus M(q)$, $d(p, x) \leq 2r^*$, and similarly, for all $x \in C_j \cup M(q) \setminus M(p)$, $d(q, x) \leq 2r^*$. So, when p and q serve as centers of $C_i \cup M(p) \setminus M(q)$ and $C_j \cup M(q) \setminus M(p)$, respectively, the cost of clustering is at most $2r^*$. But this partition differs from \mathcal{OPT} by $|M| = \epsilon n + 1$ points, so it is not ϵ -close to \mathcal{OPT} . This contradicts $(2, \epsilon)$ -approximation stability. \square

D Proofs from Section 4

D.1 Symmetric k -center

Lemma 14. *Given a clustering instance satisfying $(3, \epsilon)$ -perturbation resilience and all optimal clusters are size $> 2\epsilon n$. Assume there are two points from different clusters which are $\leq r^*$ apart from each other. Then there exists a partition $S_x \cup S_y$ of S such that for all $p, q \in S_x$, $d(p, q) \leq 2r^*$ and similarly for all $p, q \in S_y$, $d(p, q) \leq 2r^*$.*

Proof. First we prove the lemma assuming that a CCC2 exists, and then we prove the other case. When a CCC2 exists, we do not need the assumption that two points from different clusters are close.

Case 1: There exists a CCC2. If there exists a CCC, then denote c_x as a CCC for C_y . If there does not exist a CCC, then denote c_x as a CCC2 for C_y . We will show that all points are close to either C_x or C_y . c_x is distance $\leq r^*$ to all but ϵn points in C_y . Therefore, $d(c_x, c_y) \leq 2r^*$ and so c_x is distance $\leq 3r^*$ to all points in C_y . Consider the following d' .

$$d'(s, t) = \begin{cases} \min(3r^*, 3d(s, t)) & \text{if } s = c_x, t \in C_y \\ 3d(s, t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation because $d(c_x, C_y) \leq 3r^*$. Then by Lemma 3, the optimal score is $3r^*$. Given any $p \in S$, the set of centers $\{c_l\}_{l=1}^k \setminus \{c_y\} \cup \{p\}$ achieves the optimal score, since c_x is distance $3r^*$ from C_y , and all other clusters have the same center as in \mathcal{OPT} (achieving radius $3r^*$). Therefore, this set of centers must create a partition that is ϵ -close to \mathcal{OPT} , or else there would be a contradiction. Then from Fact 12, one of the centers in $\{c_l\}_{l=1}^k \setminus \{c_y\} \cup \{p\}$ must be the center for the majority of points in C_y under d' . If this center is c_l , $l \neq x, y$, then for the majority of points $q \in C_y$, $d(c_l, q) \leq r^*$ and $d(c_l, q) < d(c_z, q)$ for all $z \neq l, y$. Then by definition, c_l is a CCC for C_y . But then l must equal x , so we have a contradiction. Note that if some c_l has for the majority of $q \in C_y$, $d(c_l, q) \leq d(c_z, q)$ (non-strict inequality) for all $z \neq l, y$, then there is another equally good partition in which c_l is not the center for the majority of points in C_y , so we still obtain a contradiction. Therefore, either p or c_x must be the center for the majority of points in C_y under d' .

If c_x is the center for the majority of points in C_y , then p must be the center for the majority of points in C_x (it cannot be a different center c_l , since c_x is a better center for C_x than c_l by definition). Therefore, each $p \in S$ is distance $\leq r^*$ to all but ϵn points in either C_x or C_y .

Now partition all the non-centers into two sets S_x and S_y , such that $S_x = \{p \mid \text{for the majority of points } q \in C_x, d(p, q) \leq r^*\}$ and $S_y = \{p \mid p \notin S_x \text{ and for the majority of points } q \in C_y, d(p, q) \leq r^*\}$.

Then given $p, q \in S_x$, there exists an $s \in C_x$ such that $d(p, q) \leq d(p, s) + d(s, q) \leq 2r^*$ (since both points are close to more than half of points in C_x). Similarly, any two points $p, q \in S_y$ are $\leq 2r^*$ apart. See Figure 2b. This proves case 1.

Case 2: There does not exist a CCC2. Now we use the assumption that there exist $p \in C_x, q \in C_y, x \neq y$, such that $d(p, q) \leq r^*$. Then by the triangle inequality, p is distance $\leq 3r^*$ to all points in C_x and C_y . Consider the following d' .

$$d'(s, t) = \begin{cases} \min(3r^*, 3d(s, t)) & \text{if } s = p, t \in C_x \cup C_y \\ 3d(s, t) & \text{otherwise.} \end{cases}$$

This is a 3-perturbation because $d(p, C_x \cup C_y) \leq 3r^*$. Then by Lemma 3, the optimal score is $3r^*$. Given any $s \in S$, the set of centers $\{c_l\}_{l=1}^k \setminus \{c_x, c_y\} \cup \{p, s\}$ achieves the optimal score, since p is distance $3r^*$ from $C_x \cup C_y$, and all other clusters have the same center as in \mathcal{OPT} (achieving radius $3r^*$). Therefore, this set of centers must create a partition that is ϵ -close to \mathcal{OPT} , or else there would be a contradiction. Then from Fact 12, one of the centers in $\{c_l\}_{l=1}^k \setminus \{c_x, c_y\} \cup \{p, s\}$ must be the center for the majority of points in C_x under d' .

If this center is $c_l, l \neq x, y$, then for the majority of points $t \in C_x, d(c_l, t) \leq r^*$ and $d(c_l, t) < d(c_z, t)$ for all $z \neq l, x, y$. Then by definition, c_l is a CCC2 for C_x , and we have a contradiction.

Similar logic applies to the center for the majority of points in C_y . Therefore, p and s must be the centers for C_x and C_y . Since s was an arbitrary noncenter, all noncenters are distance $\leq r^*$ to all but ϵn points in either C_x or C_y .

Now partition all the non-centers into two sets S_x and S_y , such that $S_x = \{p \mid \text{for the majority of points } q \in C_x, d(p, q) \leq r^*\}$ and $S_y = \{p \mid p \notin S_x \text{ and for the majority of points } q \in C_y, d(p, q) \leq r^*\}$.

Then given $p, q \in S_x$, there exists an $s \in C_x$ such that $d(p, q) \leq d(p, s) + d(s, q) \leq 2r^*$ (since both points are close to more than half of points in C_x). Similarly, any two points $p, q \in S_y$ are $\leq 2r^*$ apart. This proves case 2. \square

Lemma 15. *Given a set U of k elements and a set C (disjoint from U) of $k + 2$ elements, and each $u \in U$ ranks all elements in C without ties. It is not possible that for all sets $C' \subseteq C, |C'| = k$, each $c \in C'$ is ranked highest by exactly one $u \in U$.*

Proof. We will prove this by induction, starting at $k = 3$. (In fact, the lemma can be proven directly, but it is notationally less taxing to prove the main part of the lemma for $k = 3$.) Given $u_1, u_2, u_3 \in U, c_1, c_2, c_3, c_4, c_5 \in C$, and preference lists for each $u \in U$, such that for all $C' \subseteq C, |C'| = 3$, each $c \in C'$ is ranked highest by exactly one $u \in U$. Call this the *unique ranking property*.

Without loss of generality, say that c_1, c_2 , and c_3 are ranked highest by u_1, u_2 , and u_3 , respectively. Now consider the triple $\{c_1, c_2, c_x\}$, for $x = 4$ or $x = 5$. Since c_1 is ranked highest by u_1 and c_2 is ranked highest by u_2 , u_3 must rank c_x higher than c_1 and c_2 . Similar logic holds for the sets $\{c_1, c_3, c_x\}$, and $\{c_2, c_3, c_x\}$, and we conclude that u_1, u_2 , and u_3 each rank c_4 and c_5 second-highest with respect to c_1, c_2 , and c_3 .

Now consider the set $\{c_1, c_4, c_5\}$. u_1 ranks c_1 highest, and WLOG, let u_2 rank c_4 higher than c_5 . It follows that u_3 must rank c_5 higher than c_4 .

Case 1: u_1 ranks c_4 higher than c_5 . Then there is a contradiction in the set $\{c_3, c_4, c_5\}$ because u_1 and u_2 both rank c_4 higher than c_3 and c_5 .

Case 2: u_1 ranks c_5 higher than c_4 . Then there is a contradiction in the set $\{c_2, c_4, c_5\}$ because u_1 and u_3 both rank c_5 higher than c_2 and c_4 .

Our base case is now proven. The inductive step follows easily. Assume the unique ranking property does not hold for every $|U| = i - 1, |C| = i + 1$. Assume there exists $U = \{u_1, \dots, u_i\}, C = \{c_1, \dots, c_{i+2}\}$,

and M which satisfies the unique ranking property. As before, WLOG c_1, \dots, c_i are ranked highest by u_1, \dots, u_i , respectively. Then let $U' = U \setminus \{u_i\}$, $C' = C \setminus \{c_i\}$, and $u \in U$ has the same preference list as before, but with c_i removed. In order for U, C to satisfy the unique ranking property, it must be true that U', C' satisfy the unique ranking property. Otherwise we would be able to find a $C'' \subseteq C'$ in which there exists a $c \in C''$ not ranked highest by any $u \in U'$. Then in $C'' \cup \{c_i\}$ and U , u_i ranks c_i highest, so c will still not be ranked highest by any $u \in U$. This contradicts our inductive hypothesis. \square

D.2 Proof of Theorem 18

Theorem 18. *For all $\alpha \geq 1$ and $\epsilon > 0$, finding the optimal solution for k -center under (α, ϵ) -perturbation resilience is NP-hard.*

Proof. Given a value $\alpha \geq 1$ and $\epsilon > 0$. We show a reduction from the standard symmetric k -center problem, which is NP-hard. Given a k -center instance (S, d) with optimal partition \mathcal{OPT} , and let D denote the diameter of the dataset, i.e., $D = \max_{p, q \in S} d(p, q)$.

We construct an (α, ϵ) -perturbation resilient k' -center instance (S', d') as follows. Add all the points from S to S' , so $S \subseteq S'$. Add $N = n/\epsilon$ additional points p_1, \dots, p_N . Now we define d' : for all $p, q \in S$, $d'(p, q) = d(p, q)$. For all p_i and $q \in S'$, $d'(p_i, q) = \alpha(D + 1)$. Finally, let $k' = k + N$.

Now, in this clustering instance, note that all p_i are distance $\alpha(D + 1)$ from every other point. Therefore, to obtain a clustering with radius $< \alpha(D + 1)$, we must put each p_i in its own cluster. Then we have the points in S left to cluster, with k centers. The optimal way to cluster S is \mathcal{OPT} , and the maximum cluster radius in S is $< D$ by construction. Then clearly for any $r < D$, there exists a solution for (S, d) k -center with max radius $\leq r$ if and only if there exists a solution for (S', d') k' -center with max radius $\leq r$.

(S', d') is (α, ϵ) perturbation resilient: given an perturbation d'' of d' . Note that if the original \mathcal{OPT} of (S, d) has max radius r^* , then we can achieve a max radius of αr^* on (S', d') by keeping each p_i in its own cluster. Any perturbation in which each p_i is not in its own cluster must have max radius at least $\alpha(D + 1) > \alpha r^*$. Call the optimal partition under (S', d') , \mathcal{C} and the optimal partition under (S', d'') , \mathcal{C}' . Note $|S'| = N + n = n/\epsilon + n$. Then $\frac{|S|}{|S'|} = \frac{n}{n/\epsilon + n} = \frac{n}{n(\epsilon + 1)/\epsilon} = \frac{\epsilon}{\epsilon + 1} < \epsilon$.

By the above argument, \mathcal{C} and \mathcal{C}' must be at least ϵ -close. Therefore, (S', d') is (α, ϵ) -perturbation resilient. \square

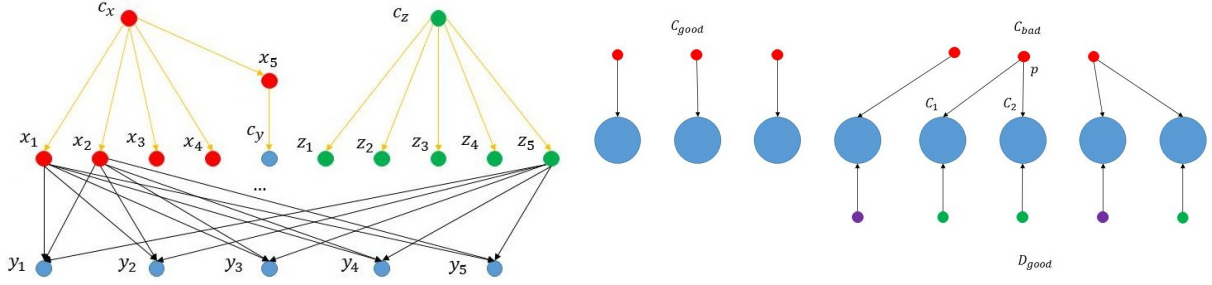
D.3 Asymmetric k -center

See Figure 4a for an example of a $(3, \epsilon)$ - perturbation resilient instance with one bad center.

Lemma 25. *The clustering instance in Figure 4a satisfies $(\alpha, 1/18)$ -perturbation resilience.*

Proof. $n = 18$, so we will argue that under any α -perturbation, at most one point switches clusters. Clearly the optimal centers are c_x, c_y , and c_z , and $r^* = 1$. Under any α -perturbation, the set of optimal centers must contain c_x and c_z , since no other points are close to x_1, \dots, x_5 and z_1, \dots, z_5 . For the final center, we cannot pick any y_i , since y_i is not close to y_j for $j \neq i$. However, we can pick any x_i or z_i , and no matter which point we pick, it will be closer to y_1, \dots, y_5 than c_x and c_z (even when distances are scaled by a factor of α). Then at most one point (the new center) switches clusters, so the instance is still $1/18$ -close to the optimal clustering. \square

Lemma 20. *Given a $(3, \epsilon)$ -perturbation resilient asymmetric k -center instance such that all optimal clusters are size $> 2\epsilon n$, then there are at most 6 centers c_i such that $\exists q \notin C_i$ and $d(q, c_i) \leq r^*$.*



(a) A $(3, \epsilon)$ -perturbation resilient asymmetric k -center instance with one bad center (c_y). The orange arrows are distance 1, and the black arrows are distance $\frac{1}{\alpha}$. (b) An illustration of Case 2. The blue balls are the optimal clusters. The purple balls are the set D' .

Figure 4: Illustrations of an instance with bad centers and the proof of Lemma 20

Proof. Assume the lemma is false. The first step is to construct a set C of $\leq k - 3$ points which are $\leq 3r^*$ from every point in S . Once we find C , we will show how to find 3 dummy centers which contradict $(3, \epsilon)$ -perturbation resilience.

By assumption, there exists a set B , $|B| \geq 7$, of centers c_i such that $\exists q \notin C_i$ and $d(q, c_i) \leq r^*$. Define $a(i)$ as the center of q 's cluster (for all $c_i \in B$). Then $d(a(i), q) \leq r^*$, and so $d(a(i), C_i) \leq d(a(i), q) + d(q, c_i) + d(c_i, C_i) \leq 3r^*$. So, one might think we can remove B from the set of optimal centers, and the remaining centers are still $3r^*$ from all points in S . However, what if $\exists c_i \in B$ such that $a(i)$ is also in B ? Then we cannot take out the entire set B . Our goal is to find $B' \subseteq B$ such that $\forall c_i \in B'$, $a(i) \notin B'$ and $|B'| \geq 3$ (this will allow us to set $C = \{c_i\}_{i=1}^k \setminus B'$).

Construct a graph $G = (V, E)$ whose vertices are $c_i \in B$. For each $c_i \in B$, if $a(i) \in B$, then add a directed edge $(c_i, a(i))$. Then every point has out-degree ≤ 1 . Finding B' corresponds to finding ≥ 3 points with no edges to one another. Consider a connected component $G' = (V', E')$ of G . Then $|E'| \geq |V'| - 1$. Since every vertex has out-degree ≤ 1 , $|E'| \leq |V'|$. Then we have two cases.

Case 1: $|E'| = |V'| - 1$. Then G' is a tree, and we can find a set of $\left\lceil \frac{|V'|}{2} \right\rceil$ vertices with no edges to one another. Case 2: $|E'| = |V'|$. Then G' contains a cycle, and we can find a set of $\left\lfloor \frac{|V'|}{2} \right\rfloor$ vertices with no edges to one another.

It follows that we can always find $\left\lceil \frac{|V|}{3} \right\rceil$ vertices with no edges to one another (equality when G consists only of disjoint 3-cycles). For $|B| \geq 7$, there exists such a set B' of size ≥ 3 . Then we have the property that $c_i \in B' \implies a(i) \notin B'$.

Now let $C = \{c_i\}_{i=1}^k \setminus B'$. By construction, B' is distance $\leq 3r^*$ to all points in S . Consider the following d' : increase all distances by a factor of 3, except $d(a(i), p)$, for i such that $c_i \in B'$ and $p \in C_i$, which we increase to $\min(3r^*, 3d(a(i), p))$. Then by Lemma 3, the optimal radius is $3r^*$. Therefore, the set C achieves the optimal score even though $|C| \leq k - 3$. Then we can pick any combination of 3 dummy centers, and they must all result in clusterings ϵ -close to \mathcal{OPT} . We will show this is not possible, and there must be a contradiction.

Recall Fact 17 from the previous section. For each cluster C_i , there exists a unique point p in S ranked first for C_i , which means that in an optimal set of centers containing p , then p will always be the center for C_i . Call this point $c(i)$. Now we define the following sets. Partition C into $C_{good} = \{c \in C \mid \exists i \text{ s.t. } c = c(i)\}$ and $C_{bad} = C \setminus C_{good}$. Then $|C| = |C_{good}| + |C_{bad}|$. Furthermore, let $D_{good} = \{c \notin C \mid \exists i \text{ s.t. } c =$

$c(i)\}$, the rest of the ‘good’ points. $|C_{good}| + |D_{good}| = k$, since there is one good point per cluster. Then $(k - |D_{good}|) + |C_{bad}| \leq k - 3 \implies |C_{bad}| + 3 \leq |D_{good}|$.

In the last section, we found a set of size $k + 2$, and any subset of size k were valid centers. But now, we have a set of C fixed points, which must always be in the set of centers. However, we make the following observation. The points in C_{good} will always be the center for the same clusters. Thus, they will never affect our analysis of what clusters the dummy centers choose; they are irrelevant to our analysis. In fact, if all points in C are good, then we can reduce to the setting in the previous section.

Case 1: $C_{good} = C$, i.e., all centers in C are irrelevant to the analysis of the dummy centers. Let $k' = k - |C|$. Then pick a set E of $k' + 2$ arbitrary points from $S \setminus C$. Given any $E' \subseteq E$ of size k' , then $C \cup E'$ must create a clustering that is ϵ -close to \mathcal{OPT} . Since C always grab the same clusters, we can use Lemma 15 to get a contradiction.

Case 2: $|C_{good}| < |C|$. We need C_{bad} to have at least 2 points to find a contradiction. If $|C_{bad}| = 1$, then add an arbitrary point $p \in S \setminus C \setminus D_{good}$ to C_{bad} . Then $|C| \leq k - 2$, and $|C_{bad}| \geq 2$, and we still have that $|C_{bad}| + 2 \leq |D_{good}|$. For all clusters C_i such that $a(i) \in D_{good}$, there exists a $c \in C_{bad}$ which is ranked highest by C_i among all other points in C_{bad} . Since $|C_{bad}| < |D_{good}|$, by the Pigeonhole Principle, there must exist some $p \in C_{bad}$ which is ranked highest in C_{bad} to two different clusters C_i , for i such that $c(i) \in D_{good}$. Call these clusters C_1 and C_2 . Then we obtain a contradiction as follows. Let the set of centers be $C \cup D'$, where $D' \subseteq D_{good}$ such that $|D'| = k - |C|$, and D' does not contain $c(1)$ or $c(2)$. See Figure 4b.

This is possible because $|D'| = k - |C| = k - |C_{good}| - |C_{bad}| = k - (k - |B_{good}|) - |C_{bad}| = |B_{good}| - |C_{bad}| \leq |B_{good}| - 2$. Every point $s \in C_{good} \cup D'$ is closest to one cluster, so if s were closest to two clusters, the solution would not be ϵ -close. And by construction, p is the closest element in C_{bad} to two different clusters, neither of which are hit by $C_{good} \cup D'$. Therefore, the clustering defined by $C \cup D'$ is not ϵ -close to \mathcal{OPT} , and we have a contradiction. \square

Theorem 19. *Algorithm 2 runs in polynomial time and outputs a clustering that is ϵ -close to \mathcal{OPT} , for $(3, \epsilon)$ -perturbation resilient asymmetric k -center instances such that all optimal clusters are size $> 2\epsilon n$.*

Proof. From Lemma 20, we know that all but $x \leq 6$ centers are in the symmetric set A .

It is possible that the symmetric 2-approximation does not return a solution of size $\leq 2r^*$ for $k' = k - x$, if there are points $p \in C_i$ such that $p \in A$ but $c_i \notin A$ (but at the very least, the algorithm will find a solution of radius $\leq 2r^*$ for $k' = k$, since A has a solution of radius $\leq r^*$). If such points p are returned as centers by the 2-approximation algorithm, it will be problematic to our analysis. However, the next step in the algorithm removes these points by brute force. Here is why the brute force step must be successful: there are $k - x$ points in C which are distance $2r^*$ from the $k - x$ centers in A , thus $3r^*$ to the corresponding clusters. There are also x points in S which are distance r^* to the final x clusters, namely their centers (and by definition, these x centers were not in $C \subseteq A$).

Finally, we explain why $|C|$ must be ϵ -close to \mathcal{OPT} . Create a 3-perturbation in which we increase all distances by 3, except for the distances from C to all points in their Voronoi tile, which we increase up to $3r^*$. Then, the optimal score is $3r^*$ by Lemma 3, and C achieves this score. Therefore, by $(3, \epsilon)$ -perturbation resilience, the Voronoi tiling of C must be ϵ -close to \mathcal{OPT} . This completes the proof. \square

E Proof of Theorem 21

Theorem 21. *Given a center-based clustering instance satisfying weak center proximity and cluster verifiability, Algorithm 3 outputs \mathcal{OPT} in polynomial time.*

Proof. To prove that the algorithm returns \mathcal{OPT} , it suffices to show that every step (b) adds an edge between two points from the same cluster.

We proceed by induction. Assume that on iteration t of the while loop in step 1, G contains no edges between two points from different clusters. Call this graph G_t . Now assume towards contradiction that in this round, the last edge added to A is $e = (p, q)$, where $p \in C_i$ and $q \in C_j$, $i \neq j$. Denote by G'_t the graph G' just before e is added to A . Let P' and Q' be the components of p and q in G'_t , respectively. (P' and Q' do not need to be subsets of C_i and C_j). WLOG, $f(P') < 0$, or else the merge would not have happened. Denote P as the connected component in G_t that contains p . Then $P \subset C_i$ by our inductive hypothesis. Furthermore, $c_i \in P'$, since $d(c_i, p) < d(p, q)$ by weak center proximity, so either c_i was already in P , or was added to P' before we added edge e . Then by definition of cluster verifiability, $f(P') < 0$ implies that $C_i \setminus P'$ is nonempty, so call it P'' . Call the component(s) in G_t corresponding to P'' by B_1, \dots, B_x . By our inductive hypothesis, for $1 \leq y \leq x$, $B_y \subset C_i$. By definition of cluster verifiability, $f(P'') < 0$, so these components must merge to a component outside of P'' . But by weak center proximity, each point in P'' is closer to c_i than to any point from another cluster. Therefore, the algorithm must add an edge between P'' and P after e is added to A , which contradicts our assumption that e was the last edge added to A .

Finally, the runtime of the algorithm is polynomial since each step involves searching through polynomially many edges. \square